多视图合作的联盟数据可视化分析

申雪锋　柯永振　姚　楠

(天津工业大学计算机科学与软件学院　天津 300387)

摘要: 【**目的**】针对当前联盟数据在进行知识发现过程中存在的问题, 设计联盟数据可视分析系统模型, 实现对历史信息的采集、挖掘和可视化分析。【**方法**】构建联盟数据可视分析系统模型, 搭建大数据平台, 验证模型的可用性。【**结果**】实验结果表明, 系统能够很好地对海量历史数据进行可视化分析, 并支持决策分析。【**局限**】目前的可视分析结果视图还可以进一步丰富。【**结论**】该系统可以对联盟的历史数据进行可视化分析, 为决策者提供科学化的数据支持。

关键词: 联盟数据　大数据　可视分析　借阅记录

分类号: TP311　G350

1　引　言

伴随着信息技术的飞速发展, 多单位联盟共享数据成为一种普遍的现象。如何更好地从联盟的海量信息中组织和挖掘出有意义的结果, 并建立一个交互的数据挖掘模型, 是现阶段数据联盟面临的一大挑战。可视化分析是一种通过交互式可视化界面辅助用户对大规模复杂数据集进行分析推理的科学与技术[1]。利用可视化分析方法可以解决信息量大、无法快速和有效交流的问题, 发现隐藏在海量资源中深层的潜在知识, 揭示结果的深层次内涵, 提高结果的可理解性和可认知性。

图书馆联盟是多单位共享数据资源的一个实例, 探索挖掘图书馆联盟数据资源的方法对联盟数据资源的建设有很好的借鉴意义, 本文以天津市图书馆联盟大数据为例证, 依靠可视分析方法, 进行知识发现, 实现决策分析和政策制定。

为了整合天津市高校图书馆的优质资源, 天津市于 2002 年开始建设天津高校联合数字图书馆。该图书馆联盟包括除天津大学和南开大学外的 17 所市属院校的 26 个图书馆。同时与南开大学、天津大学原有的 Unicorn 系统通过网关相互开放链接, 实现共享书目数据。图书馆联盟的核心任务是建立一个包含中外文书刊的联合书目共享系统, 实现成员馆之间的书目共享、技术共享, 完成图书馆工作中的包括采访、编目、流通、期刊管理、公共查询、馆际互借等自动化管理, 提高天津市高校图书馆自动化管理水平。

经过 15 年的发展, 联合图书馆积累了大量馆藏图书数据和读者借阅记录, 应用数据挖掘技术对读者借阅历史数据进行定量分析可以获知读者个性化阅读需求, 挖掘结果可以作为各图书馆文献采购决策的数据参考, 从而提高文献资源采选质量和馆藏资源利用率, 使文献采购工作更客观、科学、合理[2]。

2　相关工作

学者对于高校图书馆图书采购的相关研究主要集中于对单个图书馆的数据分析。赵迎春[3]采用灰色关联分析方法对高校图书馆各类图书的重要性进行分析评估, 综合考虑馆藏量、借阅量、高校重点学科建设以及读者的需求和评价等多种因素, 分析各因素的重要性程度和关联性, 对各类图书的重要性做出科学评估。但是文章只分析了图书的大类, 有一定的局限性。尹纪军[4]分析并研究了将神经网络用于智能图书采购的方法, 并以此为基础, 针对图书采购的行为特征, 设计了一个基于改进遗传神经网络的智能图书采购系统模型。李媛等[5]利用模糊综合评判法, 分析了借阅数

通讯作者: 申雪锋, ORCID: 0000-0001-7214-5266, E-mail: 812876188@qq.com。

据，并建立了高校图书馆文献资源采购的模糊综合评判模式，确定了各类型文献资源合理采购的额度。

大数据背景下，已经有一些学者[6-7]对图书的采购模式进行了一些探索，并且应用数据挖掘技术为高校图书馆的购书计划提供支持。迟春佳等[6]虽然讨论了数据挖掘在制订图书采购计划中应用的可行性，但并未给出具体实例。冯娜[7]虽然给出了实例，但该实例的数据基于调查问卷，结果的主观性较强。

国内对于多个图书馆的图书采购的研究还比较少，采用可视化分析方法的相关研究更少。图书分类信息是层次化数据，层次化数据的可视化一直是信息可视化领域的重要研究内容[8]。相关工作主要分为两类：采用显式表达的节点连接图和采用隐式表达的空间划分方法[9]。节点连接图将节点间的父子关系表示为节点之间的连线，这些连线能够清晰地展现节点间的上下级关系[10]；空间划分方法使用具有一定面积或体积的块表示数据中的个体节点，以树图及其变种[11-12]为代表。与节点连接图相比，空间划分方法一般将大部分空间用于叶子节点的呈现，导致非叶子节点间的上下级或相邻关系难以识别[13]。而在实际应用领域中，数据的组成不断趋于复杂，大部分数据不是仅具有单一的数据特征，而是同时具有多种数据特征[14]，对于这类复杂数据，Chen 等[15]提出采用两种或两种以上的可视化方法，以应对现有针对单一数据特征的可视化和可视分析方法不能满足对其分析的需求。

为了从不同层面对结果进行观察、认识和掌握，笔者采用多个视图从不同维度实现海量资源的有效组织和表达，设计并实现了联盟数据可视分析系统。

3 联盟数据可视分析系统模型

本文的联盟数据可视分析系统模型基于 Hadoop 平台，以 HDFS 作为海量数据存储平台，整个模型包括 Hadoop 基础架构、数据采集、数据预处理、数据分析以及数据可视化 5 个部分，如图 1 所示。

对各模块做如下说明：

(1) Hadoop 基础架构：提供 Hadoop 分布式数据(索引库、Hive 数据仓库、分析库)的操作接口、MapReduce 并行计算框架；

(2) 数据采集：根据具体需求采集相应数据；

(3) 数据预处理：完成数据的去重、降噪、特征提取等相关工作，为可视分析作数据准备；

(4) 数据分析：文本的向量化表示，对预处理后的数据进行关联分析、统计分析等实现分析功能；

(5) 数据可视化：基于 D3[16]可视化组件进行可视分析。

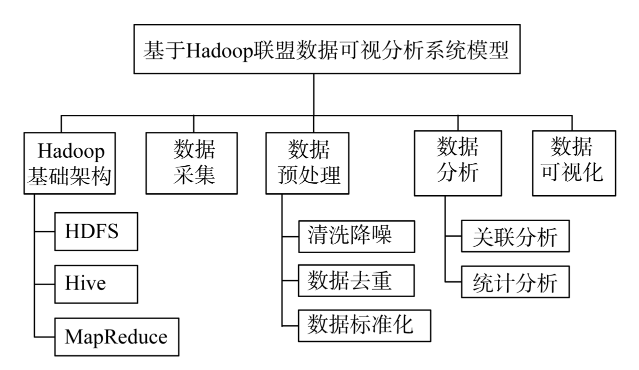


图 1　联盟数据可视分析模型

本文以图书馆联盟数据为例验证模型的可行性，主要对数据预处理模块和可视分析模块进行说明。

3.1 数据预处理

数据预处理模块需要处理借阅数据和馆藏数据。借阅数据包含 5 个字段：编号、时间、所属单位、条形码、用户名。然而由于历史原因，个别图书馆数据在合并到联合馆的过程中，数据存在以下问题：

①1907863|CJ495415|2014122615|民航大学馆|张三，此类型数据的第二项、第三项和第四项顺序错乱；

②1907864|M1214789|2014122615|商学院馆|C00624610|王五，此类型数据的第二项为冗余数据。

为了解决上述情况，需要将每条数据进行格式化，形成统一的格式，对于不符合要求的数据进行字段位置的互换或者去除。

馆藏数据的每条数据包含条形码和索书号信息。由于某一本书在书架上的排列位置会有变动的情况，所以一个条形码可能对应不同的索书号，例如条形码 ZY8027501 所对应的索书号有 D125/4、D125/1、D125/C、D125/A.L.X、D08/ELX(LS)、D751.664。从索书号集合中提取分类号需要采用分枝界限算法。

输入：索书号集合 S

输出：分类号

集合 $S(s_1,s_2,s_3\cdots s_n)$，遍历集合取 $s_i(0\leqslant i\leqslant n)$ 的首位组成集合 $K(k_1,k_2,k_3\cdots k_n)$；

遍历集合 K，计算 K 中每个元素的权重 w_j，选取 $\max w_j$ 所对应的 k_j，从 S 中选出首位以 k_j 为开头的元素集合得到 S_1；

对 S_1 的第二位进行步骤一选出集合 S_2；

依次对集合 S 中的元素的第 i 位进行运算得到 S_k，当集合 S_k 里的元素的个数唯一时结束。

3.2 数据可视化

针对具有层次化、多维特征的馆藏数据与借阅数据，笔者设计了三个视图以显示图书的层次结构信息和图书的多维属性信息。

(1) 馆藏图书显示视图

在表示馆藏图书间的层次化关系时选择节点连接图作为馆藏图书显示的视图。在树形结构中为了表示不同类别之间的数量对比关系，选择权重树作为结果的展示。所谓权重树，就是在展示节点的同时展示各个节点的权重，以节点的大小作为节点的权重大小。这里选用开源的权重树组件 Vizuly[17]作为馆藏图书的层次化结构的显示视图。

(2) 借阅图书与馆藏图书的对比视图

为了分析某类图书的使用率，定义图书借阅比 r_i 如公式(1)所示。

$$r_i = lent_i / stock_i \tag{1}$$

其中，i 代表某类图书，$stock_i$ 代表该类图书馆藏量，$lent_i$ 代表该类图书借阅量。为了表示不同图书类别在数量上差异，笔者设计了第二个视图，该视图中每一类图书由一段圆弧和两个三角形表达出馆藏数量、借阅量和借阅比三个维度的信息。使用 D3 的弧生成器作为基本图形的框架，每个弧都包含该弧段的起始角度、终止角度、内半径以及外半径等信息，使用多边形布局算法完成该视图。多边形布局算法如下。

输入：弧序列 $arc\{a_j\}_{j=1}^n$，数据 $\{d_j\}_{j=1}^n$，参数 λ,α,β

输出：弧序列 $arc\{b_j\}_{j=1}^n$，弧的宽度 $w_{j,i}$，三角形的顶点 $(x_{j,i,m}\big|_{m=1}^6, y_{j,i,m}\big|_{m=1}^6)$

使用弧生成器生成弧序列 $arc\{a_j\}_{j=1}^n$；

对于弧 a_j，计算每段弧的起始角度 $s_{j,i}$、终止角度 $e_{j,i}$；

根据参数 λ,α,β 以及起始角度 $s_{j,i}$、终止角度 $e_{j,i}$，得出弧的宽度 $w_{j,i}$，三角形的顶点 $(x_{j,i,m}\big|_{m=1}^6, y_{j,i,m}\big|_{m=1}^6)$；

依据弧序列的大小终止，输出 $arc\{b_j\}_{j=1}^n$，$w_{j,i}$，$(x_{j,i,m}\big|_{m=1}^6, y_{j,i,m}\big|_{m=1}^6)$。

本文中多边形布局算法的参数 $\lambda = 100$，参数 $\alpha = \beta = 2$。此外该算法还需要其他参数包括：外半径 R，内半径 r，圆弧的宽度 w，数据 d，圆弧 arc，圆弧的起始角度 s，圆弧的终止角度 e。

圆弧的宽度计算如公式(2)所示。

$$w_i = (R - r) \times (1 - (d_i / \lambda)) + r \tag{2}$$

弧上的点横、纵坐标计算如公式(3)和公式(4)所示。

$$x_{i,m} = R \times \sin(\theta), m \in [1,6] \tag{3}$$

$$y_{i,m} = R \times \cos(\theta), m \in [1,6] \tag{4}$$

弧外的点横、纵坐标计算公式如公式(5)和公式(6)所示。

$$x_{i,m} = (R + d_i) \times \sin((\theta + (s_i + e_i)/\alpha)/\beta),\ m \in [1,6] \tag{5}$$

$$y_{i,m} = (R + d_i) \times \cos((\theta + (s_i + e_i)/\alpha)/\beta),\ m \in [1,6] \tag{6}$$

当 $m = 1,2,4$ 时，$\theta = s_i$；当 $m = 3,5,6$ 时，$\theta = e_i$。

(3) 图书与图书馆之间的视图

为了得到读者所借图书与图书馆之间的关系，笔者设计了第三个视图。由于高校联合图书馆共有 26 个图书馆，为了分析这 26 个图书馆中每个图书馆的读者借阅不同类别图书的情况以及不同类别的图书在不同图书馆的被借阅情况，所以将图书类别与图书馆之间的关系抽象为图 $G = (V, E)$，其中顶点 V 代表图书的类别和图书馆，边 E 代表图书类别与图书馆的关系。由于图书的类别是相互独立的，不同图书馆之间是相互独立的，图书类别与图书馆之间也是相互独立的，所以 G 是一个二部图。关于二部图的可视化参照了 Pasha[18]有关二部图可视化的有关组件。

4 实例分析

图书按照分类号可以分为 22 大类，本文选取工业技术类图书进行实例分析，着重分析该类图书的借阅比以及该类图书在不同图书馆被借阅的情况。

4.1 数据来源

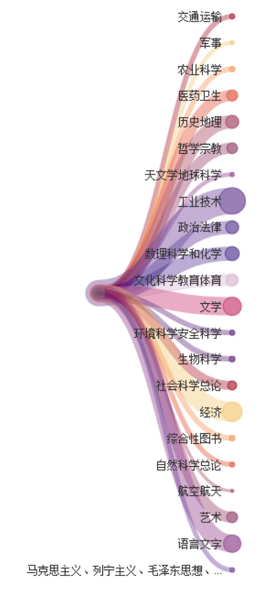
本文所采用的数据来自天津市高校数字化图书馆的 Unicorn 图书馆自动化管理系统。共有从 Unicorn 图书馆自动化管理系统开始启用截止到 2015 年 2 月 28 日的 1.6 亿条馆藏图书数据和从 2013 年 1 月 1 日到 2015 年 2 月 28 日的 396 万条借阅数据。

4.2 可视分析

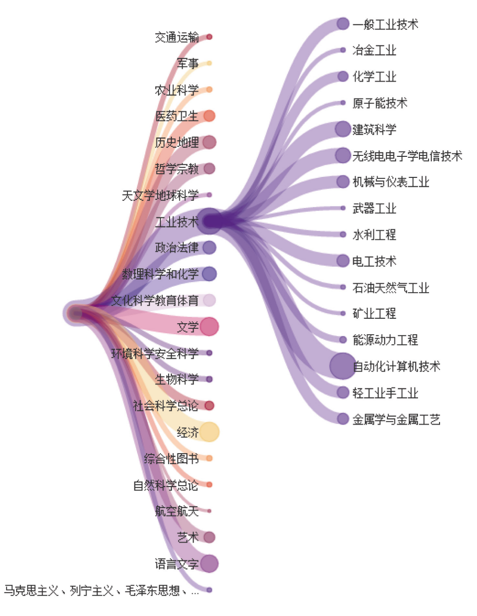
(1) 馆藏图书的分析

高校联合馆中图书的层次化结构显示如图 2 所示。由图 2(a) 可以清楚地看出馆藏图书中各类别的一个相对关系，工业技术类的图书最多，其次依次是文学

类、经济类、语言文字类，而航空航天类、天文学地球科学类的图书比较少。这种现象与17所院校中所开设的专业有关，比如只有中国民航大学开设航空航天类的专业，17所院校中没有院校开设天文学地球科学类的专业。



(a) 中图法22大类

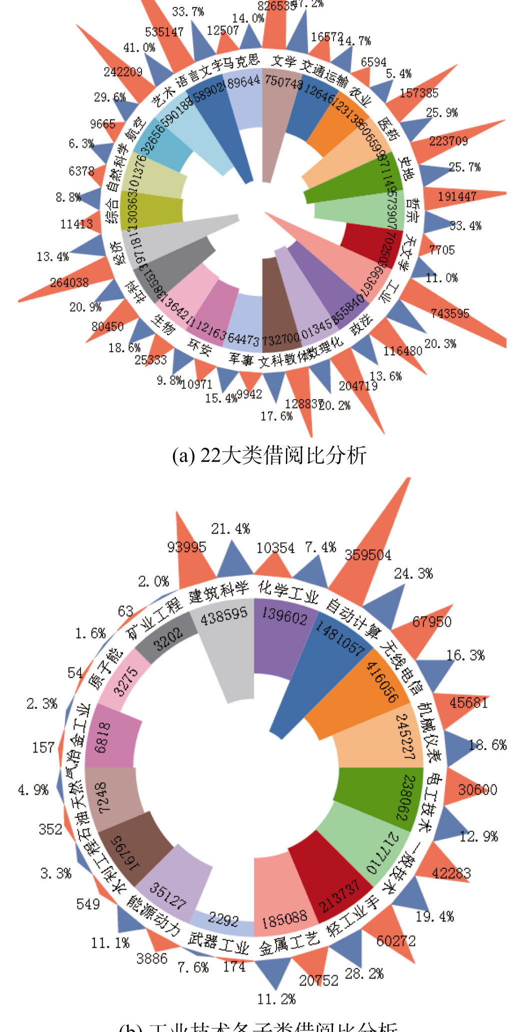


(b) 工业技术

图2　高校联合馆中图书的层次化结构

(2) 借阅图书与馆藏图书的对比分析

借阅图书与馆藏图书的对比分析如图3所示。内圆环代表每一类图书的馆藏数量，橙色三角形代表该类图书的借阅量，蓝色三角形代表借阅比 r_i。从图3(a)中可知22大类中文学类的借阅比 r_i 为47.2%，从侧面说明读者对这类图书的需求比较大。为了进一步展示22大类中各类的子类的借阅情况，选取工业技术这一类别做进一步分析，得到结果图3(b)。由图3(b)可知，自动化计算机类的图书馆藏数量最多，武器工业馆藏数量最少；自动化计算机类的借阅量最多，原子能技术类的图书借阅量最少；借阅比 r_i 最高的三类依次是轻工业手工业类(28.2%)、自动化计算机类(24.3%)、建筑科学(21.4%)，借阅比 r_i 最少的为原子能技术类(1.6%)。借阅比普遍低反应了馆藏图书的利用率比较低，这也比较符合互联网环境下用户对纸质图书需求降低的客观现实。



(a) 22大类借阅比分析

(b) 工业技术各子类借阅比分析

图3　借阅图书与馆藏图书的对比分析

(3) 图书与不同图书馆之间的分析

图 4 显示的是不同类别的图书与图书馆之间的关系视图。

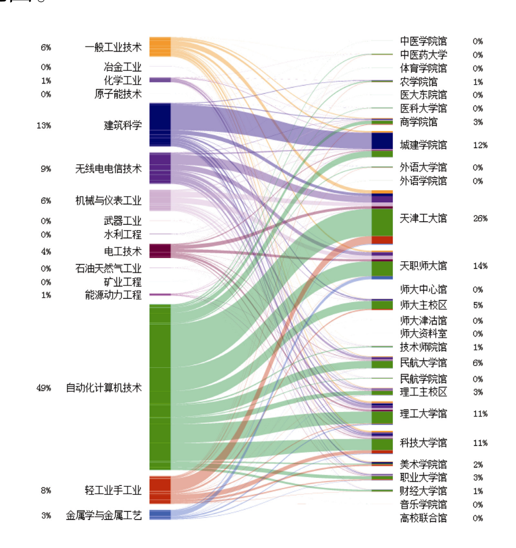
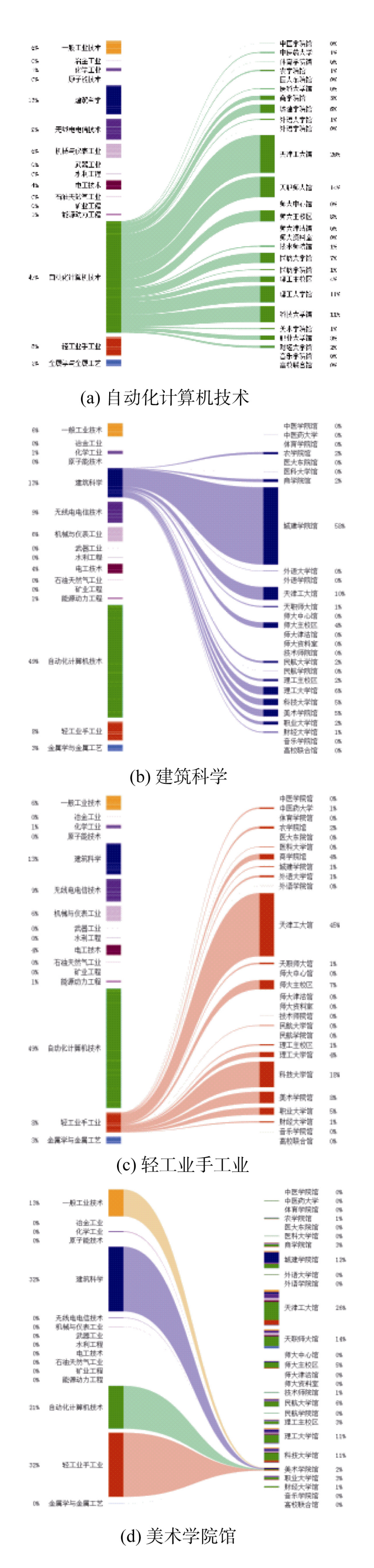


图 4　借阅图书与用户之间的分析视图

高校图书联合馆共有图书馆 26 个，从该视图中可以分析出同一类书在不同图书馆的借阅情况以及同一图书馆的读者借阅不同类别图书的情况。

对同一类书在不同图书馆的借阅情况进行分析分析如图 5 所示。由图 3 可知工业技术类的子类的借阅比 r_i 超过 20%的共有三类：自动化计算机技术类、建筑科学类、轻工业手工业类。从图 5(a)可以得到自动化计算机技术类的图书占到读者所借工业技术类图书中的 49%，读者借阅量排在前三的图书馆依次是天津工大馆(26%)、天职师大馆(14%)、城建学院馆(12%)；从图 5(b)可以看出建筑科学类的图书有 58%的被城建学院馆所借；从图 5(c)可以看出轻工业手工业的读者主要集中在天津工大馆(45%)和科技大学馆(18%)，虽然在此类别中美术学院的读者所借图书仅占 8%，但是美术学院对这类书有特别高的需求，从图 5(d)中可以看到美术学院馆的用户所借轻工业手工业类图书占到 32%，出现这种情况的原因可能和学校所设专业有关或者和读者的喜好有关，具体原因还需要进一步研究。这些数据不仅可以作为不同学校间采购经费的分配依据，还可作为本校在不同类别图书采购经费分配的依据。

进一步对借阅率比较低的三个类别进行分析如图 6 所示。



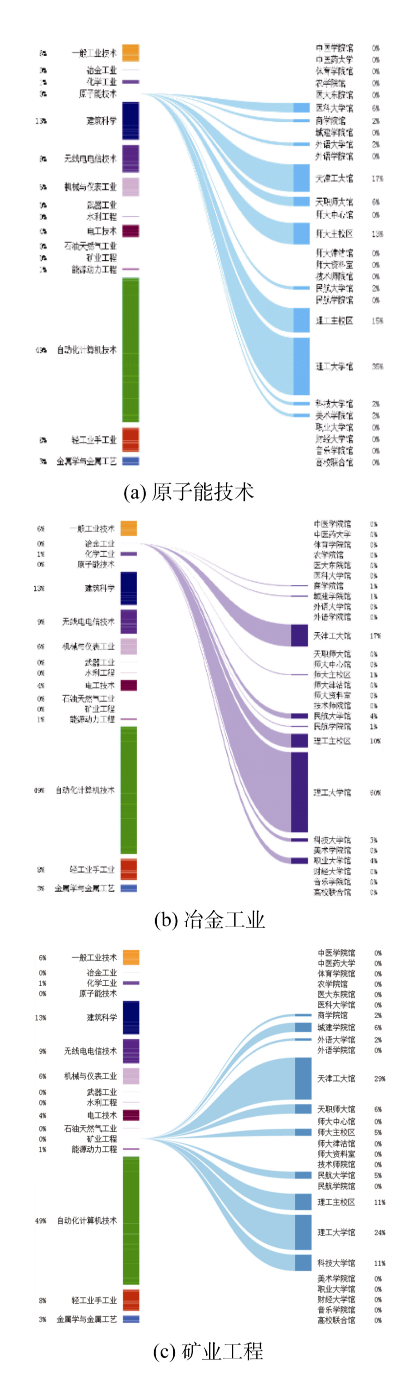
(a) 自动化计算机技术

(b) 建筑科学

(c) 轻工业手工业

(d) 美术学院馆

图 5　借阅比超过 20%的三个类别的分析



(a) 原子能技术

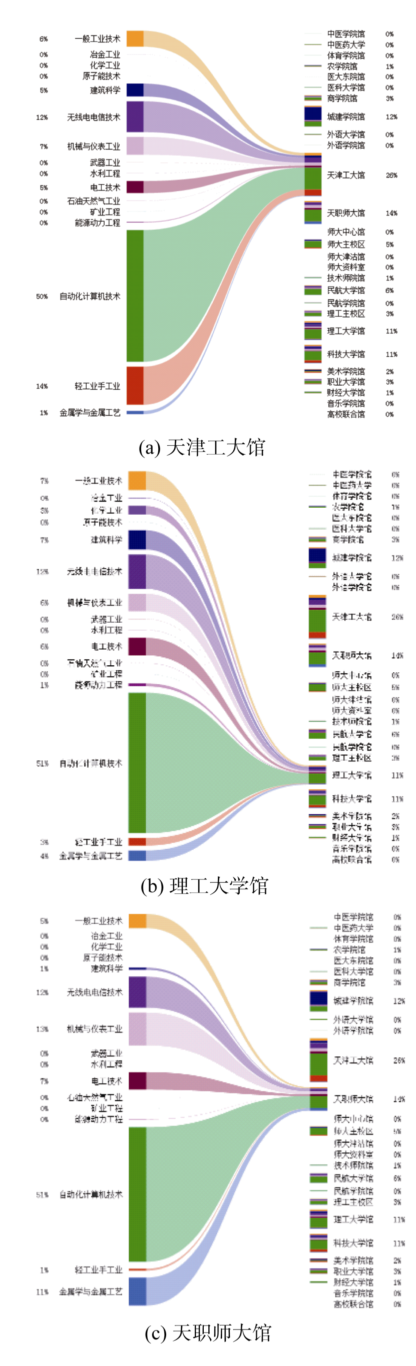
(b) 冶金工业

(c) 矿业工程

图 6 借阅率比较低的三个类别的分析

由图3(b)可知借阅比率在2%左右的共有三类：原子能技术(1.6%)、矿业工程(2.0%)、冶金工业(2.3%)。这三类图书的读者主要集中在理工大学馆、工业大学馆，但从图 6(a)可以看到医科大学馆读者借这类书的比例也达到了 6%。虽然出现这种情况的原因需要进一步研究，但是这些数据反映了读者对这一类图书的需求，采购人员可以根据读者的需求进行采购。

上面分析的是不同类别的图书与不同图书馆之间的关系，同时这个视图也可以用来分析不同图书馆的读者对于不用类别图书的不同需求，如图 7 和图 8 所示。可以发现，各个图书馆的读者普遍对工业技术类中的自动化计算机类的图书需求比较大，尤其是医学类的图书馆更是如此，采购时应关注这些图书馆读者的需求。

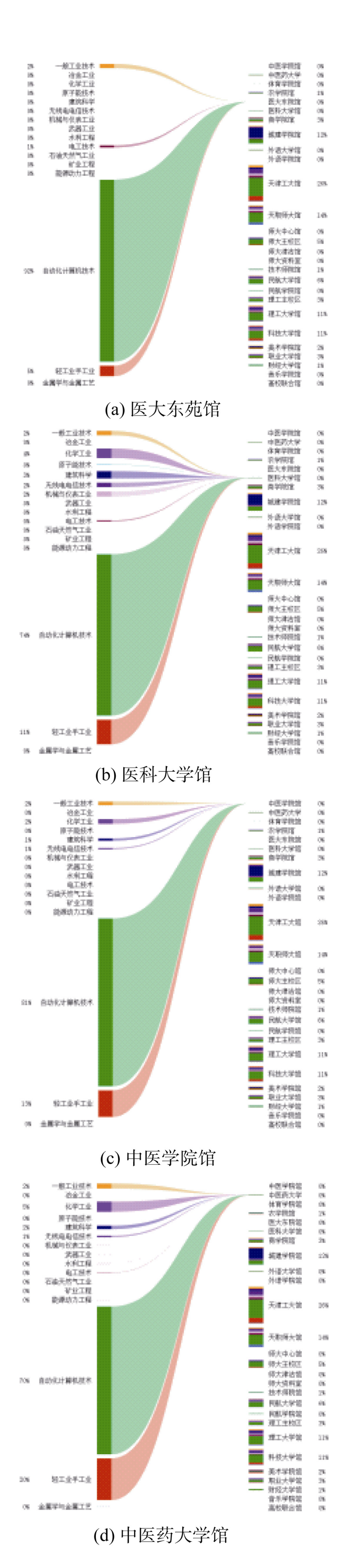


(a) 天津工大馆

(b) 理工大学馆

(c) 天职师大馆

图 7 不同图书馆的读者借阅不同类别图书的分析



(a) 医大东苑馆

(b) 医科大学馆

(c) 中医学院馆

(d) 中医药大学馆

图 8　不同图书馆的读者借阅不同类别图书的分析

5　结　语

通过对联盟图书馆的馆藏数据以及读者的借阅记录进行多维度可视化分析，笔者发现联盟图书馆存在图书借阅率较低的问题。多视图可视化分析方法除了能够更清晰地展示数据的层次结构以外，还便于进行层次下行、层次上行等自然交互，实验结果可以辅助图书馆进行图书采购活动。虽然本文只是以联盟图书馆数据进行实例分析，但提出的可视分析系统模型对其他联盟的数据依然有效，能够对联盟数据进行有效的可视化分析，发现隐藏在数据背后的潜在知识。

参考文献：

[1] 任磊，杜一，马帅，等. 大数据可视分析综述[J]. 软件学报，2014, 25(9): 1909-1936. (Ren Lei, Du Yi, Ma Shuai, et al. Visual Analytics Towards Big Data [J]. Journal of Software, 2014, 25(9): 1909-1936.)

[2] 贺德方，曾建勋. 基于语义的馆藏资源深度聚合研究[J]. 中国图书馆学报，2012, 38(200): 79-87. (He Defang, Zeng Jianxun. Study on In-depth Integration of Library Collections Based on Semantics[J]. Journal of Library Science in China. 2012, 38(200): 79-87.)

[3] 赵迎春. 灰色关联分析在高校图书馆图书采购中的应用[J]. 农业图书情报学刊，2016, 28(9): 114-118. (Zhao Yingchun. Application of Grey Relation Analysis Method in the College Libraies' Books Acquisition[J]. Journal of Library and Information Sciences in Agriculture. 2016, 28(9): 114-118.)

[4] 尹纪军. 基于改进遗传神经网络的图书采购系统研究[D]. 镇江：江苏大学，2007. (Yin Jijun. Research on Book Purchasing System Based on Improved Genetic Neural Network [D]. Zhen Jiang: Jiangsu University, 2007.)

[5] 李媛，胡蓉. 模糊综合评判法在高校图书馆文献采购中的应用[J]. 农业图书情报学刊，2014, 26(5): 72-75. (Li Yuan, Hu Rong. The Application of Fuzzy Comprehensive Evaluation Method in the Document Purchasing of University Library[J]. Journal of Library and Information Sciences in Agriculture. 2014, 26(5): 72-75.)

[6] 迟春佳，毛志勇. 基于数据挖掘的高校图书馆图书采购计划辅助决策研究[J]. 现代情报，2009, 29(7): 108-110. (Chi Chunjia, Mao Zhiyong. Research on Assistant Decision-making in Formulating University Library Book Purchasing Plan Based on Data Mining[J]. Journal of Modern Information, 2009, 29(7): 108-110.)

[7] 冯娜. 浅议基于数据挖掘的高校图书馆购书计划[J]. 农业图书情报学刊, 2016, 28(4): 112-114. (Feng Na. A Brief Discussion of University Library's Book Procurement Plan Based on Data Mining[J]. Journal of Library and Information Sciences in Agriculture, 2016, 28(4): 112-114.)

[8] 赵海森, 吕琳, 薄志涛. 面向层次化数据的变分圆形树图[J]. 软件学报, 2016, 27(5): 1103-1113. (Zhao Haisen, Lǔ Lin, Bo Zhitao. Variational Circular Treemaps for Hierarchical Data[J]. Journal of Software, 2016, 27(5): 1103-1113.)

[9] Schulz H J. Treevis. net: A Tree Visualization Reference[J]. IEEE Computer Graphics and Applications, 2011. 31(6): 11-15.

[10] Schulz H J, Schumann H. Visualizing Graphs—A Generalized View[C]//Proceedings of the Conference on Information Visualization (IV 2006). Washington, USA: IEEE Computer Society, 2006, 166-173.

[11] Tak S, Cockburn A. Enhanced Spatial Stability with Hilbert and Moore Treemaps[J]. IEEE Transactions on Visualization and Computer, 2013. 19(1): 141-148.

[12] Lam H C, Dinov I D. Hyperbolic Wheel: A Novel Hyperbolic Space Graph Viewer for Hierarchical Information Content[J]. ISRN Computer Graphics, 2012(6): 487-493.

[13] Ham F V, Wijk J V. Beamtrees: Compact Visualization of Large Hierarchies[J]. Information Visualization. 2003. 2(1): 31-39.

[14] 陈谊, 甄远刚, 胡海云, 等. 一种层次结构中多维属性的可视化方法[J]. 软件学报, 2016, 27(5): 1091-1102. (Chen Yi, Zhen Yuangang, Hu Haiyun, et al. Visualization Technique for Multi-Attrbute in Hierarchical Structure[J]. Journal of Software, 2016, 27(5): 1091-1102.)

[15] Chen Y, Zhang X Y, Feng Y C, et al. Sunburst with Ordered Nodes Based on Hierarchical Clustering: A Visual Analyzing Method for Associated Hierarchical Pesticide Residue Data[J]. Journal of Visualization, 2015. 18(2): 237-254.

[16] Bring Data to Life with SVG, Canvas and HTML[EB/OL]. [2016-11-04]. https://github.com/d3/d3.

[17] Vizuly. Weighted Tree [EB/OL]. [2016-11-04]. http://vizuly.io/product/weighted-tree/?demo=d3js.

[18] NPasha. Bipartite Graph [EB/OL]. [2016-11-04]. http://bl.ocks.org/NPasha.

作者贡献声明:

姚楠: 提供原始数据, 进行基础数据分析;
申雪锋, 柯永振: 提出研究思路, 设计研究方案, 论文最终版本修订;
申雪锋: 进行实验, 采集、清洗和分析数据, 论文起草。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 812876188@qq.com。
[1] 申雪锋. library_book, library_lent. 馆藏原始数据和图书借阅记录原始数据.
[2] 申雪锋. book_denoised. 馆藏原始数据去重之后的数据.
[3] 申雪锋. lent_Statistical data. 借阅记录统计数据.

收稿日期: 2016-11-14
收修改稿日期: 2017-02-23

Visualization of Coalition Data Based on Multi View Cooperation

Shen Xuefeng　Ke Yongzhen　Yao Nan
(School of Computer Science & Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China)

Abstract: [**Objective**] This paper proposes a data visualization model to retrieve, analyze and present historical records from a data coalition, aiming to improve the knowledge discovery. [**Methods**] We constructed a model for the visual data analysis system, and then used a big data platform to examine its feasibility. [**Results**] The proposed system could analyze massive historical data and then support the decision making procedures. [**Limitations**] The current visual analysis result views could be further improved by adding more chart templates. [**Conclusions**] The proposed system could analyze historical data from the library alliance and provide valuable information for decision makers.

Keywords: Coalition Data　Big Data　Visibility Analysis　Borrowed Records